



POLITECNICO DI TORINO Repository ISTITUZIONALE

Nonlinear Adaptive Filtering to Forecast Air Pollution

Original

Nonlinear Adaptive Filtering to Forecast Air Pollution / Mesin L.; Orione F.; Pasero E.. - STAMPA. - (2011), pp. 343-364.

Availability:

This version is available at: 11583/2461230 since:

Publisher:

InTech Open Access

Published

DOI:

Terms of use:

openAccess

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Nonlinear Adaptive Filtering to Forecast Air Pollution

Luca Mesin, Fiammetta Orione and Eros Pasero
Department of Electronics, Politecnico di Torino
Italy

1. Introduction

Air pollution is an important problem concerning the quality of living and health conditions of the population in urban areas (Pasero & Mesin, 2010). Indeed, the issue of air quality is now a major concern for many governments worldwide.

Since the early 1970s, the EU has been working to improve air quality by controlling emissions of harmful substances into the atmosphere, by improving fuel quality and by integrating environmental protection requirements into the transport and energy sectors. As the result of EU legislation, much progress has been made in tackling air pollutants. However air quality continues to cause problems. As an example, photochemical smog, particularly active in sunny days, regularly exceeds a safe limit over main European metropolitan areas. EU legislation established an hourly average of $180\mu\text{g}/\text{m}^3$ as the threshold of safe limit for ozone (Directive 2002/3/EC) beyond which authorities have to inform population.

According to the indications of the Environment Commissioner Janez Potočnik (Potočnik, 2010), air pollution is responsible for 370.000 premature deaths in EU each year. Airborne particles (e.g. Particulate Matter with diameter lower or equal to $10\mu\text{m}$, PM_{10}) are mainly present in pollutant emissions from industry, traffic and domestic heating. They can cause asthma, cardiovascular problems, lung cancer and premature death. The EU Directive 2008/50/EC requires Member States to ensure that certain limit values for PM_{10} are met. These limits, which were to be met by 2005, impose both an annual average concentration value ($40\mu\text{g}/\text{m}^3$), and a daily concentration value ($50\mu\text{g}/\text{m}^3$) which must not be exceeded more than 35 times per calendar year (European Environmental Bureau, 2005). The World Health Organization pointed out that USA traffic fatalities are over 40.000 per year, while air pollution claims 70.000 lives annually (Air Quality Guidelines, WHO, 2006). Experimental studies carried on Oslo urban site pointed out the temporal connection between cardiopulmonary inflammatory state of sick citizens and high concentration of PM in the area, especially in winter and early spring time (Schwarze P.E. et al, 2010). The impact of air pollution is on human health and on ecosystems also, as marine, freshwater, grassland, heat or forest ecosystem. The air pollution is directly linked to acidification of forests and water ecosystems, and eutrophication of soils and waters, leading to limited supply of oxygen in rivers and lakes. The benefits of air pollution abatement in the heat and grassland ecosystem, typical European farm environment, have been estimated as two hundred millions of euro each year saved (DeSmet et al, 2007). Indeed, air pollution not only affects

health care systems, as we cited above, but even local economy in terms of costs of medication, absences from work, and child and old people care expenses (Brown et al., 2002).

European environmental regulations demand the implementation of automatic procedures to prevent principal air pollutants concentration from exceeding specific alarm-thresholds in urban or suburban areas. Technology assists decision and action processes in controlling air pollution problem, implementing monitoring and forecasting techniques or automatic operating procedures in order to prevent the risk for the principal air pollutants to be above alarm thresholds. Increased efforts are made by governmental authorities for benchmarking approaches in data exchange and multi-model capabilities for air quality forecast and (near) real-time information systems, allowing information exchange between meteorological services, environmental agencies, and international initiatives.

2. Air pollution components

There are two main sources contributing to the formation of atmospheric particulate. One source can be attributed to natural phenomena such as soil erosion, volcanic eruption, presence of sea salt above oceans and other particles above water expanses. The other source is anthropogenic. Human activities impacting on the atmosphere, either directly (as primary emissions) or indirectly (as secondary emissions), bring unexpected change into the environmental processes of the ecosystem. The alteration of these natural processes is not easily absorbed by the ecosystem, bringing serious damages on it and adverse consequences on human health.

Pollutants are the elements that change the energy balance of the ecosystem. Depending on whether pollutants are introduced in the atmosphere, they can be considered primary or secondary. Primary pollutants are directly introduced in the air. They usually react with solar radiation and with other elements already in the atmosphere, changing their normal molecular structure. Secondary pollutants are not directly emitted as such, but are produced when primary pollutants react in the atmosphere (see the official site of EPA, United States Environmental Protection Agency, www.epa.gov).

From the indications of EPA, primary pollutant class includes carbon monoxides (CO), sulphur dioxides (SO₂), nitrogen oxides (NO_x), and hydrocarbons (Volatile Organic Compounds, VOCs). CO is a colorless and odorless gas, toxic for human health. CO is mainly discharged in the atmosphere by petrol fumes, as well as from steelworks and refineries, whose energy processes don't achieve complete carbon combustion. This pollutant is particularly dangerous because of its higher affinity for hemoglobin than oxygen. This could determine hypoxia in CO poisoning. High levels of CO generally occur in areas with heavy traffic congestion. SO₂ is a colorless, sour gas, which affects human respiratory system. The largest sources of SO₂ emissions are from fossil fuel combustion at power plants (73%) and other industrial facilities (20%). NO_x form quickly from volcanic and thunderstorm activity and from emission by vehicles and power plants. They are also used in fertilizers-manufacturing processes, because they can improve yield, stimulating the action of pre-existing nitrates in the ground. VOCs are organic compounds. They are divided into methane and not methane category. VOCs are not toxic, but they cause long-term chronic health effect, as reduction in visual or audible senses. Anthropogenic sources may be paints and coating, chlorofluorocarbons and fossil fuels combustion. These air pollutants are usually trapped near the ground beneath a layer of warm air, especially

under atmospheric stable conditions. Atmospheric stability occurs when the air near the surface is unable to rise, too heavy compared to the layer right above. It usually happens under high pressure atmospheric conditions. The high pressure above traps the pollutants in stagnant air surface layer (Sokhi, 2008).

Particulate matter (PM) is a complex mixture of extremely small particles and liquid droplets. Particle pollution is made up of a number of components, including acids (such as nitrates and sulfates), organic chemicals, metals, and soil or dust particles. PM can be primary or secondary (Perkins, 1974). Such powders are constituted from various pollutants, as lead, nickel, copper, cadmium and asbestos. They are easily inhaled and, depending on their dimension, they can reach and intoxicate various levels of breathing apparatus, down to alveolus, where the oxygen enhancement of hemoglobin occurs.

Ozone (O_3) is naturally present in the stratosphere, portion of the atmosphere around 12000 up to 45000 m asl, where it acts as a filter to block hazardous ultraviolet radiations. O_3 density is highest at around 25000 m asl in the ozonosphere, the low stratosphere. Thus the troposphere is not affected from the whole solar radiation load that impacts on the ozonosphere. The mechanism for high altitudes ozone formation was suggested by Sidney Chapman in Thirties (Chapman, 1932). Though the concentration of tropospheric ozone is controlled by stratosphere-troposphere exchange, in situ photochemical ozone formation plays an increasing role, especially in urban and industrial areas. At ground-level O_3 is a secondary pollutant. It is linked with adverse effects on the respiratory system (bronchitis, allergic asthma, irritations up to pulmonary edemas) and irritates mucous (Bard, 2010). At ground level ozone is created as a byproduct of the oxidation of carbon monoxide (CO) and hydrocarbons. These chemicals are called ozone precursors, and are often emitted simultaneously in the atmosphere via vehicle exhaust, industrial emissions, and other man-made sources. Photochemical ozone is catalyzed in the presence of sunlight by nitrogen oxides ($NO_x = NO + NO_2$) (Lelieveld, 2000) (Perkins, 1974). The process is called photolytic cycle for the oxides of nitrogen and causes photochemical ozone synthesis. It is summarized in the following system:



First equation of the system describes the absorption of solar radiation from NO_2 , absorption that modifies the visible wavelength of the compound. Thus the gas appears reddish-brown color over metropolitan areas, for example, in sunny, weather-steady days. Atmospheric stable conditions arise higher tropospheric O_3 density, because the high pressure traps the pollutants in stagnant air surface layer (Sokhi, 2008), inhibiting the environmental dispersion by means of wind advection and air turbulence (Geller, 2001). Second equation shows the extremely reactive O, which can react with hydrocarbon in order to have it involved in reactions that fuel nitrogen synthesis. M is any non reactive species that can absorb chemicals released in the reaction, in order to stabilize O_3 . The M body will probably be N_2 or O_2 , since these are the most available elements in the air. Third equation takes the process back to the beginning with a renewed NO_2 .

In a NO_x rich environment, the net reactions describing the oxidation of CO and VOC into respectively carbon dioxide (CO_2) and water vapor (plus carbonyl compounds), can be described as follows:



where CARB indicates Carbonyl Compounds.

In the chain of reactions resulting in (2) and (3), peroxy radicals (HO_2) are produced, which in turn oxidize nitric oxide (NO) to give nitrogen dioxide (NO_2). NO_2 is then photolyzed to give highly reactive atomic oxygen, and through reaction with oxygen a molecule of ozone (3).

3. Air pollution prediction models

Some mathematical models have been developed to simulate the time evolution of the concentration of air pollutants (Hass et al, 1995), determined by the atmospheric advection (the air transport due to wind) and by occurring chemical reactions. These models intend to simulate always more thoroughly the physical, chemical and dynamical processes which control the emission, transport and deposition of atmospheric trace species, in order to forecast their concentrations. The interaction between meteorological variables and pollutants concentration in the air is described. Reciprocal influence depends especially from wind advection and from wet and dry deposition. However sophisticated they are, these models are difficult to develop because of the wide variety of phenomena that have to be properly described for producing meaningful results, and the need of skilled personnel for setting up the experiments. These models have to take into account also the presence of urban areas, not only environmental variables, when providing local concentration values to be compared with safety thresholds.

A typical air quality physical based model is the European Air Pollution Dispersion (EURAD, http://www.eurad.uni-koeln.de/index_e.html), developed by the Rhenish Institute of Environmental Research of Germany. The EURAD system is able to simulate the evolution of pollutant tracers' in the atmosphere, from local to continental scale, by employing a nesting technique. It involves many atmospheric and chemical dedicated submodels. Physical laws that rule deterministic system models are described through systems of partial differential equations. The model can provide concentrations of pollutants of interest in the troposphere (portion of atmosphere from ground surface up to 10 km) over Europe and their transfer from air to soil or water ecosystems. Chaotic systems, as environmental one, are usually described as deterministic, so if their initial state was known exactly, their future state can be predicted. However, the precision with which the initial state can be studied strongly affects the capacity of modeling its future state (see Section 6.1).

A quicker and effective way to produce reliable outcomes of future concentrations of air pollutants can be achieved by employing black-box models for time-series forecasting (Sjöberg et al, 1994). These models are able to learn from environmental time-series, as meteorological and air pollution-concentrations datasets, the deterministic rules governing the process. No information regarding the complex physical and chemical dynamics involved in the process is used to develop such models. These methods allow real time and low cost local numerical predictions based on the analysis of extensive datasets. Black-box models perform a mathematical mapping between independent and dependent variables in a data set, by employing parameters which are not related to the physics of the complex system under study. These models can "learn" the underlying relationship between the

variables from past recordings, and they can produce forecasts of future output values accordingly. A model with specific topology and a finite number of free parameters is first chosen so that a large set of input-output relationships can be represented by proper choices of the parameters. As an example, Artificial Neural Networks (ANN) satisfy the universal approximation property, which means that they can approximate a nonlinear map as precisely as needed by increasing the number of parameters. Once chosen the model, an optimization algorithm is used to select the parameters in order to fit as close as possible to the data.

The major requirement for employing black box models for urban air quality forecasting is the availability of sufficiently long and accurate time-series for all the input and output variables involved in the study. This entails that when meteorological data are employed as input for predicting future concentrations of pollutants, the related observations should be recorded as close as possible to the air quality monitoring station, and with a similar sampling frequency.

Compared to complex physically based models, black box methods are faster, simpler and make use of a minor number of assumptions. Essentially, black box methods are based on the statistical assumption that conditions similar to those to be predicted can be found in a database of past measurements used to train the model (the training set). This requires the training set to be large enough that all specific events, including rare ones, are represented. Deterministic and black box models are not completely disjointed, even if their implementation is very different. The very fact that a certain predicted output can match a specific system condition is a deterministic hypothesis assumed by both techniques.

4. Determinism and stationarity

Air pollution data are affected by many time varying processes, like daily and seasonal atmospheric effects and further emissions. This poses the problem of stationarity. A time-series is formally defined as stationary if joint probabilities of finding it in certain conditions are independent of time (Papoulis, 1984). This means that a set of time-series should be recorded in the same conditions to define joint probabilities. Nevertheless, usually ergodicity of the series is assumed, and ensemble averages are substituted by time averages. Even under such an assumption, when we are working with finite series, inferring joint probabilities is a challenging task. It is in fact needed that all possible phenomena pertaining to the dynamics of the system can be properly represented in the time-series at disposal.

Another approach to discuss stationarity concerns the properties of the system from which the series is selected. Suppose that the environment can be described as a set of complicated, unknown deterministic rules. We could describe the dynamics of such a complex system with a vector field acting on an unknown number of state variables:

$$\frac{d\vec{x}}{dt} = \vec{F}(\vec{x}, t) \quad (4)$$

where \vec{x} is the vector of state variables and \vec{F} is a set of functions defining the vector field. A time-series, as the concentration of a pollutant sampled over time, can be considered as a measure extracted from the system. Such a measure can be modeled by an unknown function of the state variables:

$$y(t) = g(\tilde{x}(t)) \quad (5)$$

The system is said to be autonomous if the vector field does not change in time, which means:

$$\tilde{F} = \tilde{F}(\tilde{x}) \quad (6)$$

If the system is autonomous, we can expect that a sort of stationarity characterizes also the time-series that can be selected by making a measure. Nevertheless, we cannot expect that the statistical properties of an arbitrary measure are neither constant in time nor that they have only periodic trends. Indeed, the complex system determining the evolution of the pollutant of interest is nonlinear and can show chaotic behaviour (Kantz & Schreiber, 1997). Moreover, when working with finite series, we have again problems in understanding if the series can be assumed as extracted from a stationary, autonomous deterministic system or not. Indeed, we can say that the concept of stationarity for a finite series depends on the considered time period (Kantz & Schreiber, 1997) and on its relation with the time constants of the considered system. If the investigated time period is short enough that the system did not manifest any trend or evolution within it, the series appears to be stationary. On the other hand, if the system is affected by a trend which has a significant effect on the time period considered, the time-series can not be considered stationary even if it was indeed extracted from an autonomous system. In such a case, extending the period of investigation, the system could manifest a deterministic variation (as in the case of seasonal effects) demonstrating a stable stationary dynamics (called ciclo-stationary, in case of periodic variations). Thus, we could say that, in order to consider stationary a time-series, it should be measured over a time range which is much longer than the longest characteristic time scale which is relevant for the evolution of the system.

If the system has a deterministic flow which is constant in time (i.e., it is autonomous), the time-series extracted from it have dynamics determined by deterministic rules. In such a case we could, at least in principle, use past events to learn the dynamics of the considered time-series. Assuming that the space of possible events is densely covered by the trajectory $y(t)$ used to train the model, we could build a good predictor (if measurement noise is negligible) of future events by exploring what happened in the past in situations similar to the investigated one.

Sometimes we are interested in studying a time-series extracted from a system which is not stationary at all. This is the case in which we are trying to predict air pollution dynamics close to a place in which a new polluting plant is installed, for example. Another instance is the case in which a system for air pollution prediction is trained in a place and then it is moved in a nearby region. A further example could be the variation of the data introduced by a calibration of the sensor in which an offset is eliminated by adding a constant. In all these cases the dynamics of the environment from which the air pollution data are extracted or the measuring apparatus undergo an abrupt change. Thus, the prediction system is required to *adapt* to the new data, as limited information can be found on the training examples contained in the database (Widrow & Winter, 1988).

5. Spatial and temporal adaptation

From the discussion above, we conclude that adaptation can be useful at least for two reasons.

First, when we need a data-driven model to perform reliable predictions on time-series measured from an environment which is stationary. Data-driven models adapt to the specific data set or to the specific environment from which the air pollution data are recorded. The prediction model can be trained on past measurements and free parameters can be fixed, minimising a cost function which penalizes the prediction mistakes on the training data. The same methodology can be used to fit the same model to different data extracted from different environments. Nevertheless, the optimal choice of the parameters of the model will be different when considering data collected from different places. We could say that the forecast tool built in this way is adaptive in the sense that it is able to adapt to and it is determined by the specific data to be processed. We'll use the term "spatial adaptation" to indicate this property of the filter.

Second, when we need an adaptive filtering approach to perform tried and true predictions on data recorded from a non stationary environment. In such a case we need a filter with parameters that are not fixed, but can vary when the environment undergoes a change. The parameters of the model will be changed dynamically reducing a cost function, which is not built using only training data but also using the new acquired data (as soon as they are available). In this way the filter can learn new dynamics of the system, which were not represented in the training data. We can define this second feature of the adaptive filter as "temporal adaptation", as the filter is not fixed in time, but is defined by parameters which are dynamically changed to adapt to temporal variations of the data.

6. Nonlinear local adaptive prediction methods

6.1 Introduction

Many works in the literature (Foxall et al. 2001, Marra et al 2003) indicate that the air pollution time-series are nonlinear. Let us assume that the air pollution series is a measure taken from a system that can be described with a vector flow, as shown by equations (4) and (5). An autonomous, nonlinear system with more than two state variables can show chaotic behavior (Strogatz, 1994). This is surely the case for air pollution applications. This implies that similar past events could evolve into very different conditions. Thus, we expect that the dynamics of the system from which the time-series was extracted is sensitive to small differences amongst initial conditions. Such small differences can be amplified by the nonlinear dynamics toward very different final conditions.

The Lyapunov exponent is a measure of the rate of exponential divergence of trajectories starting from neighboring points (Kantz & Schreiber, 1997). The prediction horizon is usually very short and related to the inverse of the Lyapunov exponent (Haykin, 1999). Indeed, initial conditions are known with some uncertainty due to additive noise, always present on experimental, real data. The inaccuracy of measuring the initial conditions is amplified exponentially for a chaotic system, determining a larger and larger imprecision on predicted values.

Nonlinear dynamics could be poorly approximated using linear methods, so that nonlinear models must be built. Both meteorological and air pollution data are useful to adapt the model and learn how to predict pollutants concentration. The following steps are in general used to develop an adaptive local prediction tool (Pasero & Mesin, 2010):

- a. measurement of information on the investigated environment through specific sensors;
- b. pre-processing of raw time-series data (to reduce noise content or to extract optimal features);

- c. selection of a model representing the dynamics of the investigated process;
- d. choice of optimal parameters of the model in order to minimize a cost function measuring the error in forecasting the data of interest; the mean square error is usually chosen as cost function;
- e. validation of the prediction, which possibly guides the selection of an alternative model.

The predicted concentration is the result of a nonlinear processing of the possibly linearly filtered input data.

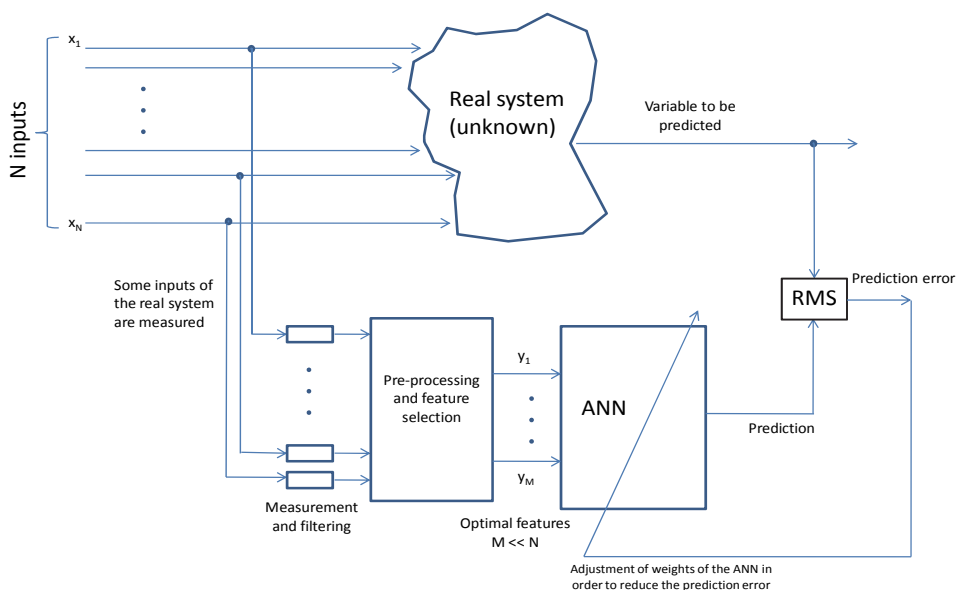


Fig. 1. Sketchy representation of the training of the nonlinear filter. The details of the real system are neglected. Some of the state variables of the real system are measured and linearly filtered in order to reduce noise or to remove undesired trends. The measured variables are then processed (e.g. by PCA or ICA) and essential features are selected. The ANN processes such features and produces a prediction which is compared with the actual value of the variable. The prediction error guides the update of the weights (i.e. free parameters of the ANN) in order to optimize the prediction.

In this chapter we focus on ANNs, which have been often used as a prognostic tool for air pollution (Perez et al., 2000; Božnar et al., 2004; Cecchetti et al., 2004; Slini et al., 2006; Karatzas et al., 2008). We discuss in the following an application of ANN to the regression and system identification analysis in order to predict the dynamics of local ozone concentration. Ozone is an important pollutant due to its extensive occurrence in EU territory (not only metropolitan) and many efforts have been played up to provide appreciable data sets.

A nonlinear filter based on an ANN is designed updating adaptively the weights in order to minimize the prediction error (Figure 1). A short introduction of methods for features

selection is discussed below, together with an elementary description of the properties of ANNs and methods for training them and for system identification.

6.2 Features selection

The selection of optimal features that are going to be used as the input of the forecasting tool is of paramount importance. Selecting good features is useful to facilitate data visualization and understanding, to reduce the measurement and storage requirements, to bring down the noise content, to defy the curse of dimensionality, to improve prediction or classification performance.

Simple methods to select the best input variables are based on adding one feature at a time and exploring the approximation error or starting with a complex network with many input variables and pruning it removing the variables with lower influence. For example, a pruning algorithm is described in (Corani, 2005) to develop an air pollution predictor model. A model considered large enough to capture the desired input-output relationship was initially tested. Then a measure of the contribution of each variable to the network efficiency is computed each iteration. A pruning algorithm removes the less influential feature from the network architecture. The procedure continues iteratively until just a unique input data remains in the network. The optimal brain surgeon (OBS) algorithm was used to determine the variable to be eliminated, based on estimate of the increase of error on the training set resulting from the removal of each input variable from the network architecture. The network providing the minimum error on a validation set is finally chosen.

Input features may also be pre-processed before being applied to the network. Linear filters may be used in order to remove noise or a trend from the input data. In order to remove redundant information, but still preserving the main energy or information contained in the data, the input variables can be linearly combined using Principal Component Analysis (PCA) or Independent Component Analysis (ICA). The components obtained by such linear combinations can be ranked in order of the energy or of the amount of information they preserve of the input variables. The components preserving the maximum energy or information are usually selected, whereas the others are neglected, obtaining data compression. Moreover, in case in which the neglected components have high noise content, such a data compression allows also improving the signal to noise ratio.

Specifically, PCA determines the amount of redundancy in the measured data by cross-correlation and estimates a linear transformation W , which reduces this redundancy to a minimum. The first principal component is the direction of maximum variance in the data. The other components can be obtained iteratively searching for the directions of maximum variance in the space of data orthogonal to the subspace spanned by already determined principal directions. As an alternative, principal components and their energy content can be obtained by solving the eigenvalue problem for the cross-correlation matrix:

$$\hat{\mathbf{R}}_{xx} = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix} \quad (7)$$

where r_{ij} is the correlation between the i^{th} and the j^{th} data (Pasero & Mesin, 2010; Mesin et al., 2011).

ICA extracts from input data some features which are statistically independent, preserving the total information content and minimizing the mutual information. ICA performs a linear

transformation between the data and the features to be determined. Central limit theorem guarantees that a linear combination of variables has a distribution that is “closer” to a Gaussian than that of any individual variable. Assuming that the features to be estimated are independent and non-Gaussian (but possibly one of them), the independent components can be determined by applying to the data the linear transformation that maps them into features with distribution which is as far as possible from Gaussian. Thus a measure of non-Gaussianity is used as an objective function to be maximized by a given numerical optimization technique with respect to possible linear transformations of the input data. Different methods have been developed considering different measures of Gaussianity. The most popular methods are based on measuring kurtosis, negentropy or mutual information (Hyvarinen, 1999; Mesin et al., 2011).

Another interesting algorithm was proposed in (Koller and Sahami, 1996). The mutual information of the features is minimized (in line with ICA approach), using a backward elimination procedure where at each state the feature which can be best approximated by the others is eliminated iteratively (see Pasero & Mesin, 2010 for an air pollution application of this method). Thus in this case the mutual information of the input data is explored, but there is no transformation of them (as done instead by ICA).

A further method based on mutual information is that of looking for the optimal input set for modelling a certain system selecting the variables providing maximal information on the output. Thus, in this case the information that the input data have on the output is explored, and features are again selected without being transformed or linearly combined. However, selecting the input variables in term of their mutual information with the output raises a major redundancy issue. To overcome this problem, an algorithm was developed in (Sharma, 2000) to account for the interdependencies between candidate variables exploiting the concept of Partial Mutual Information (PMI). It represents the information between a considered variable and the output that is not contained in the already selected features. The variables with maximal PMI with the output are iteratively chosen (Mesin et al, 2010).

Many of the methods indicated above for feature selections are based on statistical processing of the data, requiring the estimation of probability density functions from samples. Different methods have been proposed to estimate the probability density function (characterizing a population), based on observed data (which is a random sample extracted from the population). Parametric methods are based on a model of density function which is fit to the data by selecting optimal values of its parameters. Other (not parametric) methods are based on a rescaled histogram. Kernel density estimation or Parzen method (Parzen, 1962; Costa et al., 2003) was proposed as a sort of a smooth histogram.

A short introduction to feature selection and probability density estimation is discussed in (Pasero & Mesin, 2010).

6.3 ANN

Our approach exploits ANNs to map the unknown input-output relation in order to provide an optimal prediction in the least mean squared (LMS) sense (Haykin, 1999). ANNs are biologically inspired models consisting of a network of interconnections between neurons, which are the basic computational units. A single neuron processes multiple inputs and produces an output which is the result of the application of an activation function (usually nonlinear) to a linear combination of the inputs:

$$y_i = \varphi_i \left(\sum_{j=1}^N w_{ij} x_j + b_i \right) \quad (8)$$

where $\{x_j\}$ is the set of inputs, w_{ij} is the synaptic weight connecting the j^{th} input to the i^{th} neuron, b_i is a bias, $\varphi_i(\cdot)$ is the activation function, and y_i is the output of the i^{th} neuron considered. Fig. 2A shows a neuron. The synaptic weights w_{ij} and the bias b_i are parameters that can be changed in order to get the input-output relation of interest.

The simplest network having the universal approximation property is the feedforward ANN with a single hidden layer, shown in Fig. 2B.

The training set is a collection of pairs $\{\bar{x}_k, d_k\}$, where \bar{x}_k is an input vector and d_k is the corresponding desired output. The parameters of the network (synaptic weights and bias) are chosen optimally in order to minimize a cost function which measures the error in mapping the training input vectors to the desired outputs. Usually, the mean square error is considered as cost function:

$$E(\bar{w}, \bar{b}) = \sum_{i=1}^N (d_i - y(\bar{x}_i; \bar{w}, \bar{b}))^2 \quad (9)$$

Different optimization algorithms were investigated to train ANNs. The main problems concern the velocity of training required by the application and the need of avoiding the entrapment in a local minimum. Different cost functions have also been proposed to speed up the convergence of the optimization, to introduce a-priori information on the nonlinear map to be learned or to lower the computational and memory load. For example, in the sequential mode, the cost function is computed for each sample of the training set sequentially for each step of iteration of the optimization algorithm. This choice is usually preferred for on-line adaptive training. In such a case, the network learns the required task at the same time in which it is used by adjusting the weights in order to reduce the actual mistake and converges to the target after a certain number of iterations. On the other hand, when working in batch mode, the total cost defined on the basis of the whole training set is minimized.

An ANN is usually trained by updating its free parameters in the direction of the gradient of the cost function. The most popular algorithm is backpropagation, a gradient descent algorithm for which the weights are updated computing the gradient of the errors for the output nodes and then propagating backwards to the inner nodes. The Levenberg-Marquardt algorithm (Marquardt, 1963) was also used in this study. It is an iterative algorithm to estimate the synaptic weights and the bias in order to reduce the mean square error selecting an update direction which is between the ones of the Gauss-Newton and the steepest descent methods. The optimal update of the parameters $\bar{\delta}_{opt}$ is obtained solving the following equation:

$$(J^T J + \lambda I) \bar{\delta}_{opt} = J^T (\bar{d} - y(\bar{x}, \bar{W})) \quad \text{where} \quad J_i = \frac{\partial y(\bar{x}_i, \bar{W})}{\partial \bar{W}}, \quad \bar{W} = (\bar{w}, \bar{b}) \quad (10)$$

where λ is a regularization term called damping factor. If reduction of the square error E is rapid, a smaller damping can be used, bringing the algorithm closer to the Gauss-Newton

method, whereas if an iteration gives insufficient reduction in the residual, λ can be increased, giving a step closer to the gradient descent direction. A few more details can be found in (Pasero & Mesin, 2010).

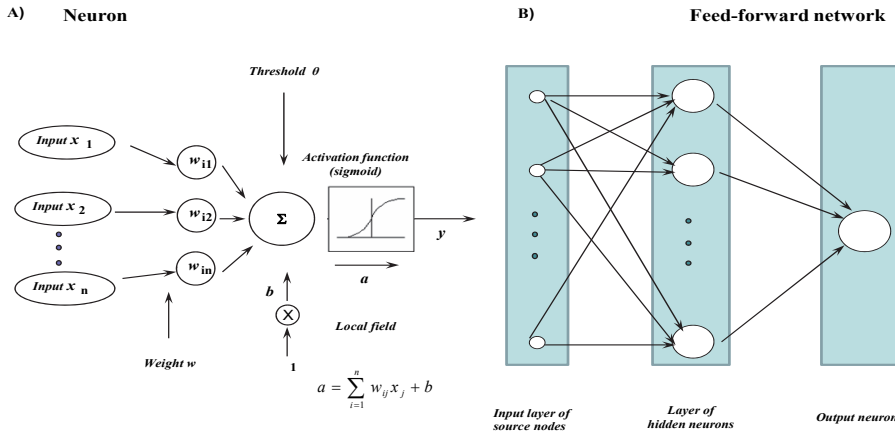


Fig. 2. A) Sketchy representation of an artificial neuron. B) Example of feedforward neural network, with a single hidden layer and a single output neuron. It is the simplest ANN topology satisfying the universal approximation property.

Due to the universal approximation property, the error in the training set can be reduced as much as needed by increasing the number of neurons. Nevertheless, it is not needed to follow also the noise, which is always present in the data and is usually unknown (even no information about its variance is assumed in the following). Thus, reducing the approximation error beyond a certain limit can be dangerous, as the ANN learns not only the determinism hidden within the data, but also the specific realization of the additive random noise contained in the training set, which is surely different from the realization of the noise in other data. We say that the ANN is overfitting the data when a number of parameters larger than those strictly needed to decode the determinism of the process are used and the adaptation is pushed so far that the noise is also mapped by the network weights. In such a condition, the ANN produces very low approximation error on the training set, but shows low accuracy when working on new realizations of the process. In such a case, we say that the ANN has poor generalization capability, as cannot generalize to new data what it learns on the training set. A similar problem is encountered when too much information is provided to the network by introducing a large number of input features. Proper selection of non redundant input variables is needed in order not to decrease generalization performance (see Section 6.2).

Different methods have been proposed to choose the correct topology of the ANN that provides a low error in the training data, but still preserving good generalization performances. In this work, we simply tested more networks with different topology (i.e., a different number of neurons in the hidden layer) on a validation set (i.e., a collection of pairs of inputs and corresponding desired responses which were not included in the training set). The network with minimum generalization error was chosen for further use.

6.4 System identification

For prediction purposes, time is introduced in the structure of the neural network. For immediately further prediction, the desired output y_n at time step n is a correct prediction of the value attained by the time-series at time $n+1$:

$$y_n = \hat{x}_{n+1} = \varphi(\bar{w} \cdot \bar{x} + b) \quad (11)$$

where the vector of regressors \bar{x} includes information available up to the time step n . Different networks can be classified on the basis of the regressors which are used. Possible regressors are the followings: past inputs, past measured outputs, past predicted outputs and past simulated outputs, obtained using past inputs only and the current model (Sjöberg et al., 1994). When only past inputs are used as regressors for a neural network model, a nonlinear generalization of a finite impulse response (FIR) filter is obtained (nonlinear FIR, NFIR). A number of delayed values of the time-series up to time step n is used together with additional data from other measures in the nonlinear autoregressive with exogenous inputs model (NARX). Regressors may also be filtered (e.g., using a FIR filter). More generally, interesting features extracted from the data using one of the methods described in Section 2 may be used. Moreover, if some of the inputs of the feedforward network consist of delayed outputs of the network itself or of internal nodes, the network is said to be recurrent. For example, if previous outputs of the network (i.e., predicted values of the time-series) are used in addition to past values of input data, the network is said to be a nonlinear output error model (NOE). Other recursive topologies have also been proposed, e.g. a connection between the hidden layer and the input (e.g. the simple recurrent networks introduced by Elman, connecting the state of the network defined by the hidden neurons to the input layer; Haykin, 1999). When the past inputs, the past outputs and the past predicted outputs are selected as regressors, the model is recursive and is said to be nonlinear autoregressive moving average with exogenous inputs (NARMAX). Another recursive model is obtained when all possible regressors are included (past inputs, past measured outputs, past predicted outputs and past simulated outputs): the model is called nonlinear Box Jenkins (NBJ).

7. Example of application

7.1 Description of the investigated environment and of the air quality monitoring station

To coordinate and improve air quality monitoring, the London Air Quality Network (LAQN) was established in 1993, which is managed by the King's College Environmental Research Group of London. Recent studies commissioned by the local government Environmental Research Group (ERG) estimated that more than 4300 deaths are caused by air pollution in the city every year, costing around £2bn a year. Air pollution persistence or dispersion is strictly connected to local weather conditions. What are typical weather conditions over London area? Precipitation and wind are typical air pollution dispersion factor. Nevertheless rainy periods don't guarantee optimal air quality, because rain only carries down air pollutants, that still remain in the cycle of the ecosystem. Stable, hot weather is typical air pollution persistence factor. From MetOffice reports we deduce rainfall is not confined in a special season. London seasons affect the intensity of rain, not the incidence. Snow is not very common in London area. It is most

likely when Arctic and Siberian winds occur from north, north-east. In the summer there are usually a few days of particularly hot weather in London. They are often followed by a thunderstorm.

In this study, we used the air quality data from the LAQN Harlington station situated in the Hillingdon borough. London Hillingdon-Harlington (LHH, 51,488 lat, -0, 416 lon) is an urban background air quality station located in Heathrow Airport zone. The station is north-east the main Heathrow runway, around 21 km west far from London City. The borough of Hillingdon is on the outskirts of the densely populated London area and its air quality is affected by the airport and road traffic, urban heating and suburb manufacturing. There are some expanses of water, small lakes, and green zones around 10 km west from LHH. The area is plain. CO, NO, NO₂ and NO_x, O₃, PM₁₀ and PM_{2.5} are the pollutants species monitored. Meteorological data was obtained by a nearby LAQN monitoring station located in the Heathrow Airport (LHA).

LHA-LHH zone should experience ozone, nitrogen oxides and carbon monoxide pollution. As we mentioned above, nitrogen oxides are in fact synthesized from urban heating, manufacturing processes and motor vehicle combustion, especially when revs are kept up, over fast-flowing roads and motorways. There are a motorway (A4) at about 2 km north from Heathrow runway and another perpendicular fast-flowing road (M4). Nitrogen oxides, especially in the form of nitrate ions, are used in fertilizers-manufacturing processes, to improve yield by stimulating the action of pre-existing nitrates in the ground. As we mentioned above, the study area is on the borderline of a green, cultivated zone west from London metropolitan area. Carbon monoxide, a primary pollutant, is directly emitted especially from exhaust fumes and from steelworks and refineries, whose energy processes don't achieve complete carbon combustion.

7.2 Neural network design and training

The study period ranged from January 2004 to December 2009, though it was reduced to only those days where all the variables employed in the analysis were available. All data considered, 725 days, were at disposal for the study and 16 predictors were selected: daily maximum and average concentration of O₃, up to three days before (6 predictors); daily maximum and average concentration for CO, NO, NO₂ and NO_x of the previous day (8 predictors); daily maximum and daily average of solar radiation of the previous day (2 predictors). Predictors have been selected according to literature (Corani, 2005; Lelieveld & Dentener, 2000), completeness of the recorded time-series, and a preliminary trial and error procedure. Efficient air pollution forecasting requires the identification of predictors from the available time-series in the database and the selection of essential features which allow obtaining optimal prediction. It is worth noticing that, by proceeding by trials and errors, the choice of including O₃ concentration up to three days before was optimal. This time range is in line with that selected in (Kocak, 2000), where a daily O₃ concentration time-series was investigated with nonlinear analysis techniques and the selected embedding dimension was 3.

Data were divided into training, validation and test set.

The training set is used to estimate the model parameters. The first 448 days and those with maximum and minimum of each selected variable were included in the training set. Different ANN topologies were considered, with number of neurons in the hidden layer

varying in the range 3 to 20. The networks were trained with the Levenberg-Marquardt algorithm in batch mode. Different numbers of iterations (between 10 and 200) were used for the training.

The validation set was used to compute the generalization error and to choose the ANN with best generalization performances. The validation data set was made of the 277 remaining days, except for 44 days. The latter represents the longest uninterrupted sequence and it has been therefore used as test dataset (see Section 7.3).

The network with best generalization performances (i.e., minimum error in the validation set) was found to have 4 hidden neurons, and it was trained for 30 iterations. Once the optimal ANN has been selected, it is employed on the test data set. The test set is used to run the chosen ANN on previously unseen data, in order to get an objective measure of its generalization performances.

Another neural network was developed from the first one, changing dynamically the weights using the new data acquired during the test. The initial weights of the adapted ANN are those of the former ANN, selected after the validation step. The adaptive procedure is performed using backpropagation batch training. For the prediction of the (n+1) observation in the data set, all the previous n-data patterns in test data set are used to update the initial weights. Also this neural network was employed on the test data set, as shown in the following section.

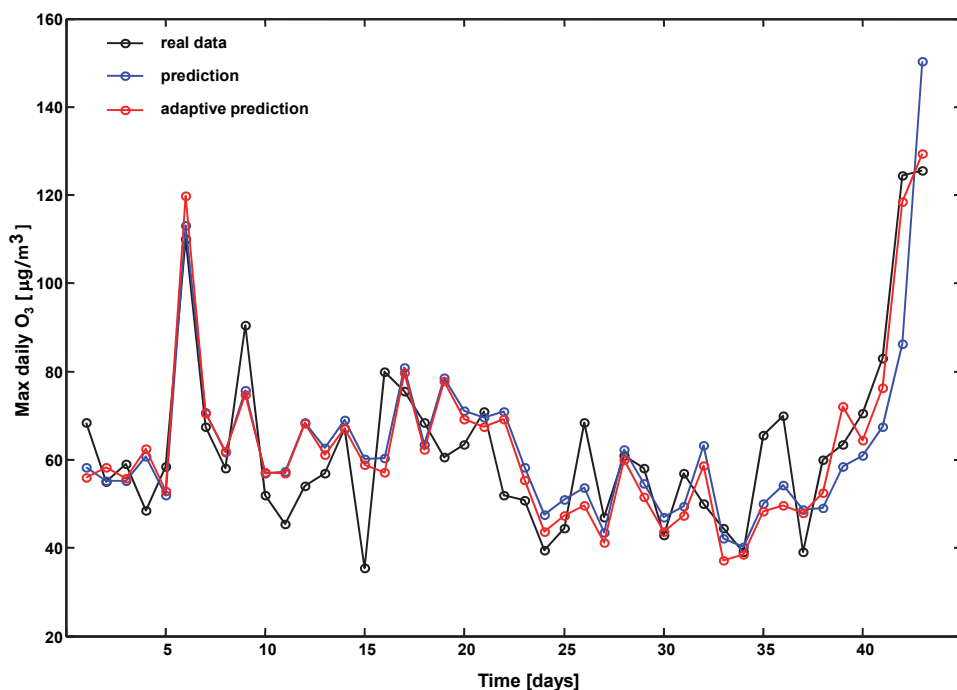


Fig. 3. Application of two ANNs to the test set.

7.3 Results

Two different ANNs are considered, as discussed in Section 7.2. The first one has weights which are fixed. This means that the network was adapted to perform well on the training set and then was applied to the test set. This requires the assumption that the system is stationary, so that no more can be learned from the new acquired data. Such an ANN is spatially adapted to the data (referring to Section 5). The second network has the same topology as the first one, but the weights are dynamically changed considering the new data which are acquired. The adaptation is obtained using backpropagation batch training, considering the data of the test set preceding the one to be predicted. Thus, temporal adaptation is used (refer to Section 5).

The results of the first ANN on the test data set are shown in Figure 3 and in Table 1 in terms of linear correlation coefficient (R^2), root mean square error (RMSE) and ratio between the RMSE and the data set standard deviation (STD). It emerges that the performances on the training and validation data set are generally good; the RMSE is below half the standard deviation of the output variable and R^2 around 0.90. A drop in the performances is noticeable on the test data set, meaning that some of the dynamics are not entirely modeled by the ANN. Performing a temporal adaptation by changing the ANN weights, a slight improvement in prediction performances is noticed as shown in Table 1. The adapted network is obtained using common backpropagation as described before. The optimal number of iterations and the adaptive step were respectively found to be 14 and 0.0019, low enough to prevent instabilities due to overtraining.

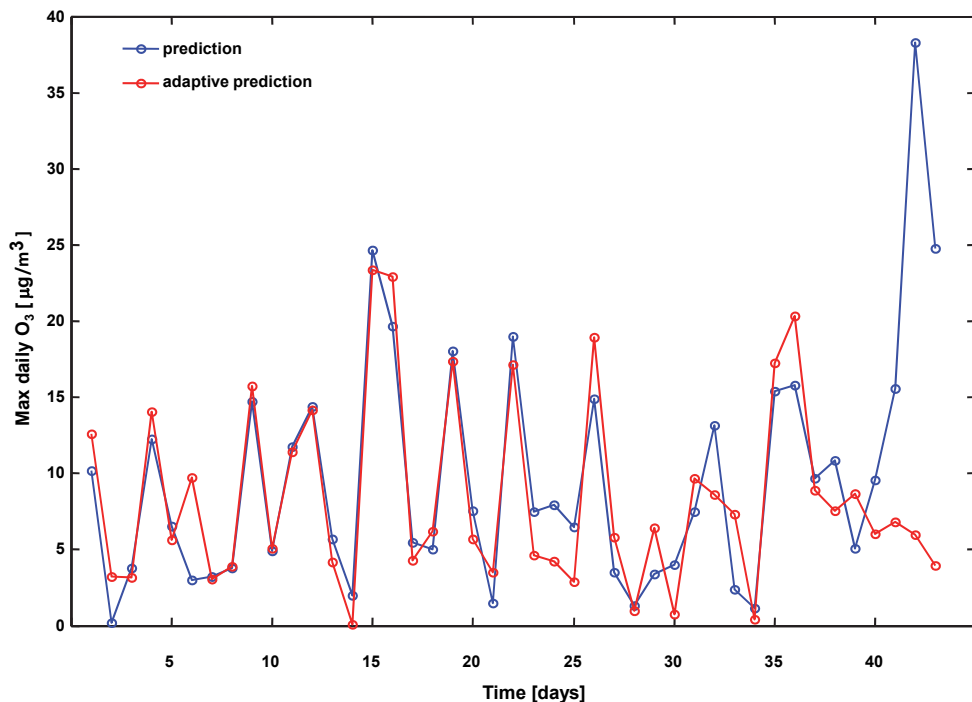


Fig. 4. Absolute error of two ANNs when applied to the test set.

DATASET	RMSE [$\mu\text{g}/\text{m}^3$]	RMSE/STD	R ²
TRAINING SET	11.19	0.45	0.89
VALIDATION SET	11.41	0.41	0.91
TEST SET (FIXED WEIGHTS)	12.35	0.62	0.79
TEST SET (TEMPORAL ADAPTATION)	10.42	0.52	0.86

Table 1. Results of application of two ANNs to the data

From the comparison of predictions in Figure 3 and most notably from the plot of the absolute errors in Figure 4, it can be seen that the adaptive network performs sensibly better towards the end of the data set, i.e. when more data is available for the adaptive training. The accuracy of the ANN model can also be compared to the performances of the persistence method, shown in Table 2. The persistence method assumes that the predicted variable at time $n+1$ is equal to its value at time n . Although very simple, this method is often employed as a benchmark for forecasting tools in the field of environmental and meteorological sciences. For example, many different nonlinear predictor models were compared to linear ones and to the persistence method in forecasting air pollution concentration in (Ibarra-Berastegi et al, 2009). Surprisingly, in many cases persistence of level was not outperformed by any other more sophisticated method. Concerning this study, however, it can be seen comparing the results in Tables 1 and 2 that the considered ANNs outperforms the persistence method in each data set considered, with improvements in terms of RMSE ranging from around 40% to 50% .

DATASET	RMSE [$\mu\text{g}/\text{m}^3$]	R ²
TRAINING SET	19.82	0.66
VALIDATION SET	20.66	0.70
TEST SET	19.85	0.43

Table 2. Results of application of the persistence method to the data

7.4 Discussion

Two predictive tools for tropospheric ozone in urban areas have been developed. The performances of the models are found to be satisfactory both in terms of absolute and relative goodness-of-fit measures, as well as in comparison with the persistence method. This entails that the choice of the exogenous predictors (CO, nitrogen oxides, and solar radiation) was appropriate for the task, though it would be interesting to assess the change in performances that can be obtained by including other reactants (VOC) involved in the formation of tropospheric ozone.

In terms of model efficiency, it has been shown that further adaptive training on the test data set may result in increased accuracy. This could indicate that the dynamics of the environment is not stationary or, more probably, that the training set was not long enough for the ANN model to learn the dynamics of the environment. However, a thorough analysis of the benefits of adaptive training can be carried out on longer uninterrupted time-

series. For instance, such a study could give insights on the optimal number of previous data patterns to be used for the adaptive steps.

Adaptive training could also be employed to improve pollutant prediction on nearby sampling stations. Since the development of air quality forecasting tools with ANNs is a data-driven process, the quantity as well as the quality of the information at disposal is of primary importance. This may severely hinder the development of accurate local models for recently installed sampling stations, or for those nodes of the monitoring network where the amount of missing/non validated data is considerable. To overcome these problems, one could first develop an ANN model for another node of the network, close enough to the one of interest and with a sufficient number of reliable data for training and validation. Once the major dynamics of the process are mapped into the ANN architecture using the former dataset, the model can be fine tuned with adaptive training to match the conditions of the chosen node, such as different reactants concentrations or local meteorological conditions.

8. Final remarks and conclusion

Many applications are not feasible to be processed with static filters with a fixed transfer function. For example, noise cancellation, when the frequency of the interference to be removed is slightly varying (e.g., power line interference in biomedical recordings), cannot be performed efficiently using a notch filter. For such problems, the filter transfer function can not be defined a-priori, but the signal itself should be used to build the filter. Thus, the filter is determined by the data: it is data-driven.

Adaptive filters are constituted by a transfer function with parameters that can be changed according to an optimization algorithm minimizing a cost function defined in terms of the data to be processed. They found many applications in signal processing and control problems like biomedical signal processing (Mesin et al., 2008), inverse modeling, equalization, echo cancellation (Widrow et al, 1993), and signal prediction (Karatzas et al, 2008; Corani, 2005).

In this chapter, a prediction application is proposed. Specifically, we performed 24-hour maximal daily ozone-concentrations forecast over London Heathrow airport (LHA) zone. Both meteorological variables and air pollutants concentration time-series were used to develop a nonlinear adaptive filter based on an artificial neural network (ANN). Different ANNs were used to model a range of nonlinear transfer functions and classical learning algorithms (backpropagation and Levenberg-Marquardt methods) were used to adapt the filter to the data in order to minimize the prediction error in the LMS sense. The optimal ANN was chosen with a cross-validation approach. In this way, the filter was adapted to the data. We indicated this process with the term "spatial adaptation". Indeed, the specific choice of network topology and weights was fit to the data detected in a specific location. If prediction is required for a nearby region, the same adaptive methodology may be applied to develop a new filter based on data recorded from the new considered region. Thus, a specific filter is adapted to the data of the specific place in which it should be used. Hence, in a sense, the filter is specific to the spatial position in which it is used. For this case, the concept of "spatial adaptation" was introduced in order to stress the difference with respect to what can be called "temporal adaptation". Indeed, once the filter is adapted to the data, two different approaches can be used to forecast new events: the transfer function of the filter could be fixed (which means that the weights of the ANN are fixed) and the prediction tool can be considered as a static filter; on the other hand, the filter could be dynamically updated considering the new

data. In the latter case, the filter has an input-output relation which is not constant in time, but it is temporally adapted exploiting the information contained in the new detected data. Both approaches have found applications in the literature. For example, in (Rusanovskyy et al. 2007), video compression coding was performed both within single frames using a “spatial adaptation” algorithm and over different frames using a “temporal adaptation” method. Both spatial and temporal adaptations were also implemented here for the representative application on ozone concentration forecast. The “spatial adaptation” of the ANN (on the basis of the training set) was sufficient to obtain prediction performances that overcome those of the persistence method when the filter was applied to the new data contained in the test set. This indicates that the training was sufficient for the filter to decode some of the determinism that relates the future ozone concentration to the already recorded meteorological and air pollution data. Moreover, applying to new data the same deterministic rules learned from the database used for training, the predictions are reliable. Nevertheless, when the filter was updated based on the new data (within the “temporal adaptation” framework), the performances were still greater. This indicates that new information was contained in the test data. The same outcome is expected in all cases in which the investigated system is not stationary or when it is stationary, but the training dataset did not span all possible dynamics.

The specific application presented in this work showed the importance of having consistent datasets in order to implement reliable tools for air quality monitoring and control. These datasets have to be filled with information from weather measurement stations (equipped with solar radiation, temperature, pressure, wind, precipitation sensors) and air quality measurement stations (equipped with a spectrometer to determine particle matters size and sensors to monitor concentration of pollutants like O_3 , NO_x , SO_2 , CO). It is important that different environmental and air pollution variables are measured over the same site, as all such variables are related by physical, deterministic laws imposing their diffusion, reaction, transport, production or removal. Indeed, local trend of air pollutants can cause air quality differences in a range of 10-20 km.

As all statistical approaches, also our filter would benefit of increasing the amount of training and test data, unavoidable condition to give the work more and more significance. Long time-series could be investigated in order to assess possible non stationarities, which temporally adapted filters could decode and counteract in the prediction process. Different sampling stations could also be investigated in order to assess the spatial heterogeneities of air pollution distribution. Moreover, the work could be extended to other consistent air pollutant datasets, in order to provide a more complete air quality analysis of the chosen site.

In conclusion, local air pollution investigation and prediction is a fertile field in which adaptive filters can play a crucial role. Indeed, data-driven approaches could provide deeper insights on pollution dynamics and precise local forecasts which could help preventing critical conditions and taking more efficient countermeasures to safeguard citizens health.

9. Acknowledgments

We are deeply indebted to Riccardo Taormina for his work in processing data and for his interesting comments and suggestions. This work was sponsored by the national project AWIS (Airport Winter Information System), funded by Piedmont Authority, Italy.

10. References

- Bard, D.; Laurent, O.; Havard, S.; Deguen, S.; Pedrono, G.; Filleul, L.; Segala, C.; Lefranc, A.; Schillinger, C.; Rivière, E. (2010). Ambient air pollution, social inequalities and asthma exacerbation in Greater Strasbourg (France) metropolitan area: the PAISA study, Artificial Neural Networks to Forecast Air Pollution, Chapter 15 of "Air Pollution", editor V. Villaniy, SCIYO Publisher, ISBN 978-953-307-143-5.
- Božnar, M.Z.; Mlakar, P.J.; Grašič, B. (2004). Neural Networks Based Ozone Forecasting. Proceeding of 9th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, June 1-4, 2004, Garmisch-Partenkirchen, Germany.
- Brown, L.R.; Fischlowitz-Roberts, B.; Larsen, J.; (2002). The Earth Policy Reader, Earth Policy Institute, ISBN 0-393-32406-0.
- Cecchetti, M.; Corani, G.; Guariso, G. (2004). Artificial Neural Networks Prediction of PM10 in the Milan Area, Proc. of IEMSs 2004, University of Osnabrück, Germany, June 14-17.
- Chapman, S. (1932). Discussion of memoirs. On a theory of upper-atmospheric ozone, Quarterly Journal of the Royal Meteorological Society, vol. 58, issue 243, pp. 11-13.
- Corani, G. (2005). Air quality prediction in Milan: neural networks, pruned neural networks and lazy learning, Ecological Modelling, Vol. 185, pp. 513-529.
- Costa, M.; Moniaci, W.; Pasero, E. (2003) INFO: an artificial neural system to forecast ice formation on the road, Proceedings of IEEE International Symposium on Computational Intelligence for Measurement Systems and Applications, pp. 216-221.
- De Smet, L.; Devoldere, K.; Vermoote, S. (2007). Valuation of air pollution ecosystem damage, acid rain, ozone, nitrogen and biodiversity - final report. Available online: http://ec.europa.eu/environment/air/pollutants/valuation/pdf/synthesis_report_final.pdf.
- Environmental Research Group, King's College London. (2010).z Air Quality project [Online]. Available: <http://www.londonair.org.uk/london/asp/default.asp>.
- European Environmental Agency EEA, (2008) Annual European Community LRTAP Convention emission inventory report 1990-2006, Technical Report 7/2008, ISSN 1725-2237.
- European Environmental Bureau EEB, (2005) Particle reduction plans in Europe, EEB Publication number 2005/014, Editor Responsible Hontelez J., December.
- European Communities, (2002) Directive 2002/3/EC of the European Parliament and of the Council of 12 February 2002 relating to ozone ambient air, Official Journal of European Community, OJ series L, pp. L67/14-L67/30. Available: <http://eur-lex.europa.eu/JOIndex.do>.
- Foxall, R.; Krcmar, I.; Cawley, G.; Dorling, S.; Mandic, D.P. (2001). Nonlinear modelling of air pollution time-series, *icassp*, Vol. 6, pp. 3505-3508, IEEE International Conference on Acoustics, Speech, and Signal Processing.
- Geller, R. J.; Dorevitch, S.; Gummin, D. (2001). Air and water pollution, Toxicology Secrets, 1st edition, L. Long et al Ed., Elsevier Health Science, pp.237-244.
- Hass, H.; Jakobs, H.J. & Memmesheimer, M. (1995). Analysis of a regional model (EURAD) near surface gas concentration predictions using observations from networks, Meteorol. Atmos. Phys. Vol. 57, pp. 173-200.

- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, Prentice Hall.
- Hyvarinen, A. (1999). Survey on Independent Component Analysis, *Neural Computing Surveys*, Vol. 2, pp. 94-128.
- Ibarra-Berastegi, G.; Saenz, J.; Ezcurra, A.; Elias, A.; Barona, A. (2009). Using Neural Networks for Short-Term Prediction of Air Pollution Levels. *International Conference on Advances in Computational Tools for Engineering Applications (ACTEA '09)*, July 15-17, Zouk Mosbeh, Lebanon.
- Kantz, H. & Schreiber, T. (1997). *Nonlinear Time-series Analysis*, Cambridge University Press.
- Karatzas, K.D.; Papadourakis, G.; Kyriakidis, I. (2008). Understanding and forecasting atmospheric quality parameters with the aid of ANNs. *Proceedings of the IJCNN*, Hong Kong, China, pp. 2580-2587, June 1-6.
- Koller, D. & Sahami, M. (1996). Toward optimal feature selection, *Proceedings of 13th International Conference on Machine Learning (ICML)*, pp. 284-292, July 1996, Bari, Italy.
- Kocak, K.; Saylan, L.; Sen, O. (2000). Nonlinear time series prediction of O₃ concentration in Istanbul. *Atmospheric Environment*, Vol. 34, pp. 1267-1271.
- Lelieveld, J.; Dentener, F.J. (2000). What controls tropospheric ozone?, *Journal of Geophysical Research*, Vol. 105, n. d3, pp. 3531-3551.
- London Air Quality Network, Environmental Research Group of King's College, London. Web page : <http://www.londonair.org.uk/london/asp/default.asp>.
- Marra, S.; Morabito, F.C. & Versaci M. (2003). Neural Networks and Cao's Method: a novel approach for air pollutants time-series forecasting, *IEEE-INNS International Joint Conference on Neural Networks*, July 20-24, Portland, Oregon.
- Marquardt, D (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics* 11: 431-441. doi:10.1137/0111030.
- Mesin, L.; Kandoor, A.K.R.; Merletti, R. (2008). Separation of propagating and non propagating components in surface EMG. *Biomedical Signal Processing and Control*, Vol. 3(2), pp. 126-137.
- Mesin, L.; Orione, F.; Taormina, R.; Pasero, E. (2010). A feature selection method for air quality forecasting, *Proceedings of the 20th International Conference on Artificial Neural Networks (ICANN)*, Thessaloniki, Greece, September 15-18.
- Mesin, L.; Holobar, A.; Merletti, R. (2011). Blind Source Separation: Application to Biomedical Signals, Chapter 15 of " *Advanced Methods of Biomedical Signal Processing*", editors S. Cerutti and C. Marchesi, Wiley-IEEE Press, ISBN: 978-0-470-42214-4.
- Met Office UK climate reports, <http://www.metoffice.gov.uk/climate/uk/averages/ukmapavge.html>.
- Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York.
- Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33: 1065-1076. doi:10.1214/aoms/1177704472.
- Pasero, E.; Mesin L. (2010). Artificial Neural Networks to Forecast Air Pollution, Chapter 10 of " *Air Pollution*", editor V. Villaniy, SCIYO Publisher, ISBN 978-953-307-143-5.

- Perez, P; Trier, A.; Reyes, J. (2000). Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment*, Vol. 34, pp. 1189-1196.
- Perkins H.C. (1974) *Air Pollution*, International Student Edition, Mc Graw Hill.
- Potočník, J. (2010).
<http://europa.eu/rapid/pressReleasesAction.do?reference=IP/10/1265&format=HTML>.
- Rusanovskyy, D.; Gabbouj, M.; Ugur, K. (2007). Spatial and Temporal Adaptation of Interpolation Filter For Low Complexity Encoding/Decoding. *IEEE 9th Workshop on Multimedia Signal Processing*, pp.163-166.
- Science Encyclopedia. <http://science.jrank.org/pages/6028/Secondary-Pollutants.html>.
- Schwarze, P.E.; Totlandsdal, A.L.; Herseth, J.L.; Holme, J.A.; Låg, M; Refsnes, M.; Øvrevik, J.; Sandberg, W.J.; Bølling, A.K.; (2010). Importance of sources and components of particulate air pollution for cardio-pulmonary inflammatory responses, Chapter 3 of "Air Pollution", editor V. Villaniy, SCIYO Publisher, ISBN 978-953-307-143-5.
- Sharma, A. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: 1 - A strategy for system predictor identification. *Journal of Hydrology*, Vol. 239, pp. 232-239.
- Sjöberg, J.; Hjalmeron, H. & L. Ljung (1994). *Neural Networks in System Identification*. Preprints 10th IFAC symposium on SYSID, Copenhagen, Denmark. Vol.2, pp. 49-71.
- Sokhi R. S. (2007), *World Atlas of Atmospheric Pollution*, Ed. Anthem Press, London.
- Slini, T.; Kaprara, A.; Karatzas, K.; Moussiopoulos, N. (2006), PM10 forecasting for Thessaloniki, Greece, *Environmental Modelling & Software*, Vol. 21(4), pp. 559-565.
- Strogatz, S.H. (1994). *Nonlinear Dynamics and Chaos*, Addison-Wesley.
- Widrow, B.; Winter, R.G. (1988), *Neural Nets for Adaptive Filtering and Adaptive Pattern Recognition*. *IEEE Computer Magazine*, Vol. 21(3), pp. 25-39.
- Widrow, B.; Lehr, M.A.; Beaufays, F.; Wan, E.; Bilello, M. (1993). *Adaptive signal processing*. *Proceedings of the World Conference on Neural Networks*, IV-548, Portland.
- World Health Organization (2006). *Air quality guidelines. Global update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide*, ISBN 92 890 2192 6.